Predicting speech intelligibility in conditions with nonlinearly processed noisy speech

Jørgensen, Søren; Dau, Torsten

Published in:
Proceedings of the International Conference on Acoustics - AIA-DAGA 2013

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Predicting speech intelligibility in conditions with nonlinearly processed noisy speech

Søren Jørgensen and Torsten Dau

Centre for Applied Hearing Research, DTU, 2800 Kgs. Lyngby, Denmark, Email: sjor@elektro.dtu.dk

Abstract

The speech-based envelope power spectrum model (sEPSM; [1]) was proposed in order to overcome the limitations of the classical speech transmission index (STI) and speech intelligibility index (SII). The sEPSM applies the signal-to-noise ratio in the envelope domain (SNR_{env}), which was demonstrated to successfully predict speech intelligibility in conditions with nonlinearly processed noisy speech, such as processing with spectral subtraction. Moreover, a multi-resolution version (mr-sEPSM) was demonstrated to account for speech intelligibility in various conditions with stationary and fluctuating interferers [2]. However, the model fails in the case of phase jitter distortion, in which the spectral structure of speech is affected but the temporal envelope is maintained. This suggests that an across audio-frequency mechanism is required to account for this distortion. It is demonstrated that a measure of the across audio-frequency variance at the output of the modulation-frequency selective process in the model is sufficient to account for the phase jitter distortion. Thus, a joint spectro-temporal modulation analysis, as proposed in [3], does not seem to be required. The results are consistent with concepts from computational auditory scene analysis and further support the hypothesis that the SNR_{env} is a powerful metric for speech intelligibility prediction.

Introduction

The most commonly used metrics for predicting speech intelligibility in noise have traditionally been the signal-to-noise ratio (SNR), as used in the Articulation Index (AI; [4]), or the modulation transfer function (MTF), as measured by the speech transmission index (STI; [5]). These metrics have been shown to account for intelligibility in a broad range of conditions with stationary noise, low-and high-pass filtering, and reverberation. However, the AI and the STI are limited to conditions with stationary interferers, due to long-term integration of the stimuli. Moreover, these metrics generally fail when noisy speech is subjected to nonlinear processing, such as noise reduction via spectral subtraction, or phase jitter distortions, arising due to variation in the supply voltages of telephone systems.

In [3], an extension to the STI was presented, denoted as the spectro-temporal modulation index (STMI), which measures the integrity of the joint spectro-temporal modulations of the speech signal. In contrast to the one-dimensional (temporal) modulation processing assumed in the STI and sEPSM, the STMI includes a two-dimensional modulation filterbank. While this more complex modulation processing stage allows the STMI to successfully account for phase jitter distortion, the decision metric is based on the (two-dimensional) MTF, similar to the STI and, therefore, it is unclear whether it can also account for spectral subtraction processing.

In contrast to STI and STMI, [1] suggested an intelligibility metric based on the signal-to-noise ratio in the envelope domain (SNR_{env}). The SNR_{env} was highly correlated with intelligibility of noisy speech processed by spectral subtraction, and consistent with the STI in conditions with reverberation. The major difference with respect to the MTF is the explicit consideration of the (intrinsic) envelope fluctuations of the noise itself, which is increased after spectral subtraction (e.g., [6]), and potentially responsible for a reduced intelligibility. The SNR_{env}-metric is calculated as part of the speech-based envelope power spectrum model (sEPSM), inspired by the EPSM [7], originally developed to account for modulation detection and masking data. However, the SNR_{env} considered in [1] was calculated from a long-term integration of the stimuli and the sEPSM must, therefore, fail in conditions with fluctuating interferers. Moreover, it is unclear whether this model can account for phase jitter distortion, since it assumes purely temporal modulation processing.

Here, the "multi-resolution" version of the sEPSM framework is considered (mr-sEPSM) [2]. The hypothesis in the framework of this model is that the SNR_{env} is increased in the dips of a fluctuating interferer and that a "short-term" estimation of the SNR_{env} will account for the intelligibility in such conditions. The model is evaluated with two categories: (i) speech mixed with a stationary interferer and two types of fluctuating interferers with very different temporal structure; and (ii) two types of nonlinearly processed noisy speech in the form of spectral subtraction and phase jitter, testing the model's ability to account for nonlinear distortion.

Model Description

The processing structure of the mr-sEPSM is illustrated in Figure 1. The first stage is a (peripheral) bandpass filterbank, consisting of 22 gammatone filters with one equivalent rectangular bandwidth [8] and third-octave spacing of their center frequencies, covering the range from 63 Hz to 8 kHz. An absolute sensitivity threshold is included such that individual gammatone filters are included only if the level of the stimulus at the output is above the absolute hearing threshold for normal-hearing listeners. The temporal envelope of each filter output is extracted via the Hilbert-transform and low-pass filtered with a cut-off frequency of 150 Hz, using a first-order Butterworth filter. The resulting envelope is analyzed by a modulation bandpass filterbank, consistsing of eight second-order bandpass filters with constant quality factor (Q=1) and octave spacing, covering the range from 2 - 256 Hz, in parallel with a third-order lowpass filter with a cut-off frequency of 1 Hz (see [1]). The running temporal output of each modulation filter is divided into short segments using rectangular windows without overlap.

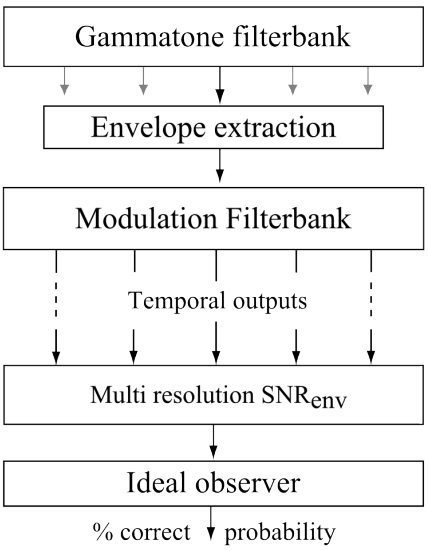

Figure 1: Schematic of the processing stages of the mr-sEPSM framework.

The duration of a segment is proportional to $1/f_{cm}$, with f_{cm} denoting the modulation filter center frequency. For example, the segment duration in the 4-Hz modulation filter is 250 ms. For each segment, the AC-coupled envelope power (variance) of the noisy speech and the noise alone are calculated separately and normalized with the corresponding long-term squared DC. The SNR_{env} of a segment is estimated from the envelope power as:

$$SNR_{env} = \frac{P_{S+N} - P_N}{P_N} \quad (1)$$

where P_{S+N} and P_N denote the envelope power of the noisy speech and the noise alone after the normalization. For each modulation filter, the running SNR_{env}-values are averaged across time, assuming that all parts of a sentence contribute equally to intelligibility. The time-averaged SNR_{env}-values from the different modulation-filters and peripheral filters are combined, using the "integration model" from [9]. The combined SNR_{env} is converted to the probability of correctly recognizing the speech item using the concept of a statistically "ideal observer" [1].

Method

Model predictions were compared to data from the literature as well as data collected for the present study. The target speech was either Danish sentences from the CLUE-speech material [10], or sentences from the TIMIT database. The data reflect either speech reception thresholds (SRTs) corresponding to the 50% point on the psychometric function or percentage of correct scores. All subjects were normal-hearing listeners.

One condition with stationary speech-shaped noise (SSN) and two conditions with fluctuating interferers were considered: (i) SSN that was amplitude modulated by an 8-Hz sinusoid (SAM), and (ii) the speech-like, but unintelligible, International Speech Test Signal (ISTS; [11]).

Moreover, five conditions with phase jitter, having different values of the jitter-factor, α [3], and six conditions with noisy speech (SSN) processed by spectral subtraction were considered. The spectral subtraction algorithm was implemented similar to [12] using six different values of the over-subtraction factor, ρ.

For the predictions, the model parameters were calibrated to a close match between the predictions and the data for the unprocessed SSN-condition. These parameters were then fixed for all other experimental conditions. Identical stimuli were used for the predictions and the measurements, except for the conditions with phase jitter where the data were obtained using sentences from the TIMIT database [3], whereas the predictions were obtained using the CLUE-sentences.

Results

Conditions with Stationary and Fluctuating Interferers

Figure 2 shows the measured SRTs (open squares) in the conditions with SSN, SAM, and ISTS interferers. The SRTs were obtained at SNRs of -3.5, -9.1, and -18 dB, respectively. The large decrease in SRT for the two fluctuating interferers reflects a clear release from speech masking. Predictions from the mr-sEPSM (filled black squares) account well for the data, while predictions from the earlier sEPSM-version [1] (filled gray symbols) clearly fail to account for the conditions with fluctuating interferers. The root mean square error (RMSE) between the measured data and the mr-sEPSM predictions amounts to 1.7 dB.

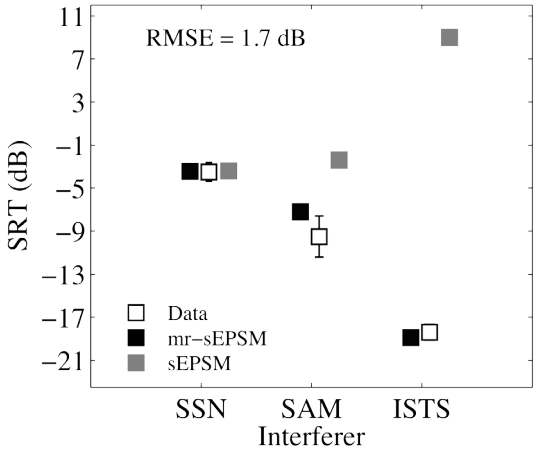

Figure 2: Results for conditions with stationary and fluctuating interferers: measured data (open squares) and predictions from the (long-term) sEPSM (gray squares) and the mr-sEPSM (black squares).

Spectral Subtraction

Figure 3 shows the data obtained by [1] (open squares) and corresponding predictions by the mr-sEPSM (filled black squares) for six conditions of ρ, where 0 denotes the reference condition with no spectral subtraction. The data show an increase of the SRT with increasing ρ, demonstrating a lower intelligibility with spectral subtraction

than without the processing. The mr-sEPSM predicts the trends in the data, although it overestimates the SRTs for ρ = 2, 4, and 8, leading to a larger RMSE (1.3 dB) compared to the sEPSM. Predictions obtained with the STMI (gray squares) suggest that the intelligibility increases after spectral subtraction, in contrast to the measured data. The STMI thus fails to account for spectral subtraction, as does the STI [1].

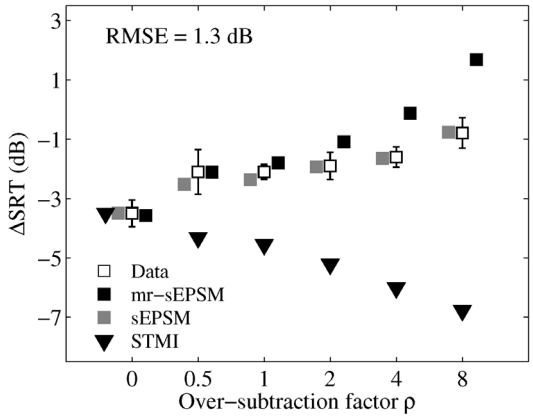

Figure 3: Results for spectral subtraction: measured SRTs (open squares) and predictions from the (long-term) sEPSM (gray squares) and the mr-sEPSM (black squares). Predictions from the STMI are indicated by filled triangles.

Phase Jitter

Figure 4 shows the measured data obtained by [3] (open circles) for speech distorted by phase jitter. The intelligibility is almost unaffected for jitter factors, α, below 0.2, but sharply decreases for values above 0.2 and remains low for higher values of α. The mr-sEPSM (filled black squares) clearly fails to account for the measured data, predicting perfect intelligibility independent of α. In contrast, the STMI (gray triangles) is in good agreement with the data.

Model Analysis

The mr-sEPSM was found to be insensitive to the phase jitter distortion. However, it is unclear whether the jitter distortion is simply not reflected in the model's internal representation, or whether the information is there, but some additional processing stage is required to capture the effect. The top-left panel of Figure 5 shows the internal representation of the mr-sEPSM in response to a sentence without phase jitter. Each trace in the panel represents the temporal output of the 4-Hz modulation filter for a subset of the peripheral filters tuned to frequencies between 0.16 and 4 kHz. The undistorted speech leads to a representation with a varying temporal structure across the different peripheral channels, reflecting the distributed speech information across the peripheral frequency channels. The top-right panel shows the same, but for the condition with α = 0.5. The jitter has clearly removed the natural variation, and increased the temporal coherence across the audio-frequency channels, reflecting a loss of speech information. Thus, the jitter distortion *is* reflected in the model's internal representation. However, an across-channel mechanism is required to capture the effect. One metric for quantifying the loss of information could be the across-audio-frequency variance of the internal representation.

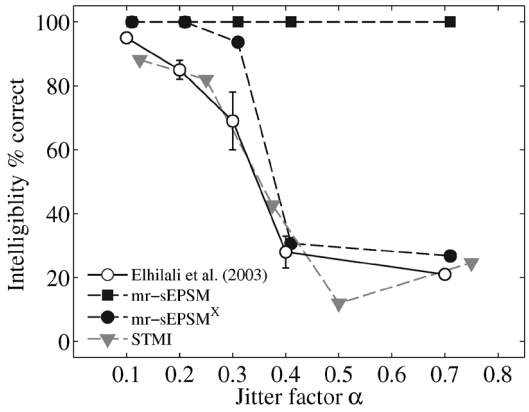

Figure 4: Result for conditions with phase jitter: measured percentage of correct response (open circles) and predictions from the mr-sEPSM (black squares), and the STMI (gray triangles). Predictions from the mr-sEPSM including across-channel processing are indicated by filed black circles.

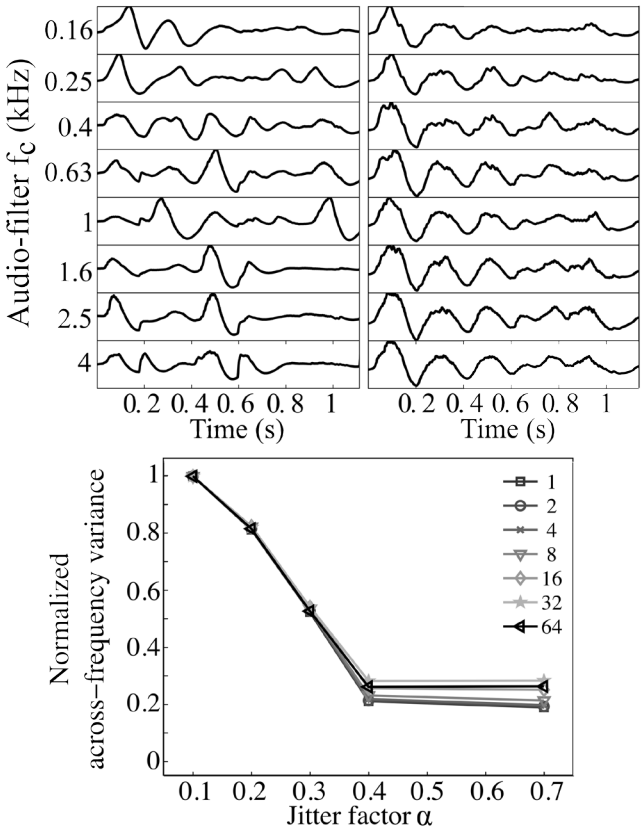

Figure 5: Top panel: the model's internal representation at the output of the 4-Hz modulation filter for a subset of the peripheral filters in response to a sentence without phase jitter (top-left panel) and with α = 0.5 (top-right panel). Bottom: The normalized across (peripheral)-frequency variance, calculated from the internal representation, as a function of the jitter factor.

The bottom panel of Figure 5 shows the normalized across-frequency variance, averaged across time, as a function of

the jitter factor for a subset of the modulation filters. The variance decreases with increasing jitter factor, in a very similar manner as the decrease of speech intelligibility shown in Figure 4. Additional predictions were obtained with a version of the model (mr-sEPSM^X), where the contribution from a given modulation filter was assumed to depend on the across-channel variance of that filter. The corresponding predictions (Fig. 4; filled black circles) largely agree with the predictions based on the STMI and the measured data.

Discussion

It was demonstrated that the multi-resolution estimation of SNR_{env} in the sEPSM-framework accounted for the speech intelligibility in conditions with a stationary and two fluctuating interferers. The critical element in estimating the running SNR_{env} was the modulation-filter dependent analysis window, ranging from 4 ms to 1000 ms, allowing the model to evaluate the increased SNR_{env} in the dips of the fluctuating interferer.

Moreover, the model accounted for the decreased intelligibility in conditions with noisy speech processed by spectral subtraction. [1] demonstrated that the SNR_{env} decreases as the spectral subtraction factor increases and that the decreasing SNR_{env} is caused by an increase of the intrinsic fluctuations in the noise-part of the noisy speech, producing more modulation masking in the model. Thus, the increased noise-fluctuations may reduce the perceptual salience of the target speech fluctuations and, thus, decrease the intelligibility for the listeners.

In contrast to the sEPSM, the STMI was shown to fail in conditions of spectral subtraction. The reason is that the STMI is based on the MTF, evaluating the difference between a clean and a processed representation, instead of the envelope signal-to-noise ratio, and therefore does not capture the effect of processing on the noise alone. However, the mr-sEPSM failed to account for phase jitter distortion, which the STMI successfully described. The model analysis demonstrated that the effect of the jitter was reflected at the output of the purely temporal modulation processing assumed in the mr-sEPSM. However, an across audio-frequency mechanism, such as the across-channel variance, was necessary to account for the perceptual data and was shown to produce similar results as the STMI. The concept of an across-channel mechanism is consistent with other recent models of comodulated masking release [13] and models of auditory scene analysis [14]. A more complex two-dimensional representation, as the one assumed in the STMI framework, may not be required.

Conclusion and Perspective

This study demonstrated that a modeling framework based on estimating the SNR_{env} in short time segments, following temporal modulation processing, can account for speech intelligibility in conditions with stationary and fluctuating interferers, as well as in conditions with noisy speech processed by spectral subtraction. Furthermore, by including an across audio-frequency mechanism, such a framework was sufficient to account for the effects of phase jitter distortion on speech intelligibility.

Including the SNR_{env} metric in more detailed models of auditory preprocessing and perception might be interesting for studying the consequences of hearing impairment on speech intelligibility.

References

[1] Jørgensen, S. and Dau, T. (**2011**). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. **130**, 1475-1487.

[2] Jørgensen, S., Ewert, S., and Dau, T. (**2013**). A multi-resolution envelope-power based model of speech intelligibility. J. Acoust. Soc. Am. (submitted).

[3] Ehilali, M., Chi, T., and Shamma, S. A. (**2003**). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun. **41**, 331-348.

[4] French, N. and Steinberg, J. (**1947**). Factors governing intelligibility of speech sounds. J Acoust Soc Am **19**, 90-119.

[5] Steeneken, H.J.M. and Houtgast, T. (**1980**) A physical method for measuring speech transmission quality. J. Acoust. Soc. Am. **67**, 318-326.

[6] Dubbelboer, F. and Houtgast, T. (**2008**). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. J. Acoust. Soc. Am. **124**, 3937-3946.

[7] Ewert, S. and Dau, T. (**2000**). Characterizing frequency selectivity for envelope fluctuations. J. Acoust. Soc. Am. **108**, 1181-1196.

[8] Glasberg, B. R. and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data", Hear. Res.. **47**, 103-138.

[9] Green, D. M. and Swets, J. A. (**1988**). Signal Detection Theory and Psychophysics, Peninsula Publishing, Los Altos California, 238-239.

[10] Nielsen, J. B. and Dau, T. (**2009**). Development of a Danish speech intelligibility test. Int. J. Audiol. **48**, 729-741.

[11] Holube, I., Fredelake, S. Vlaming, M., Kollmeier, B. (**2010**). Development and analysis of an International Speech Test Signal (ISTS). Int. J. Audiol. **49**, 891-903.

[12] Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). Enhancement of speech corrupted by acoustic noise. ICASSP **4**, 208-211.

[13] Piechowiak, T., Ewert, S. D., and Dau, T. (**2007**). Modeling comodulation masking release using an equalization-cancellation mechanism. J. Acoust. Soc. Am. **121**, 2111-2126.

[14] Ehilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (**2009**). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron. **61**, 317-329.